DOCUMENT RESUME

ED 427 352 CS 509 983

AUTHOR Hulbert-Johnson, Ruth
TITLE Evaluating Instruments for Appropriateness.
PUB DATE 1998-11-00
NOTE 19p.; Paper presented at the Annual Meeting of the National Communication Association (84th, New York, NY, November 21-24, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Communication (Thought Transfer); Ethics; Evaluation Methods; Higher Education; *Test Reliability; Test Selection; *Test Validity
IDENTIFIERS *Communication Competencies

ABSTRACT

With increasing emphasis upon accountability, assessment in general, and communication competency in particular, is placing more demands upon human and financial resources. Therefore, those making the decisions about what and how to assess various factors must make informed, ethical, and responsible decisions. To that end, this paper provides a discussion of relevant validity and reliability issues related to the selection of assessment instruments and also provides a checklist for individuals responsible for the development and/or selection of those instruments. (Contains a table of data and 19 references.) (CR)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 427 352

Evaluating Instruments for Appropriateness

Ruth Hulbert-Johnson
The Center for Excellence in Oral Communication
University of Colorado, Colorado Springs
P.O. Box 7150
Colorado Springs, CO 80933-7150
(719) 262-4775
E-mail: rjohnson@brain.uccs.edu

A paper presented at the annual convention of the
National Communication Association
New York City, November 21-24, 1998

CS509983

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

R. Hulbert-Johnson

Abstract

With increasing emphasis upon accountability, assessment in a variety of contexts is placing more demands upon human and financial resources. Therefore, those making the decisions about what and how to assess various factors must make informed, ethical and responsible decisions. To that end, this paper provides a discussion of relevant validity and reliability issues and offers suggestions for selection of assessment instruments.

INTRODUCTION

In the last 20 years assessment in general, and communication competence in particular, has received increasing attention as accountability issues move into the fore of legislator's and/or administrator's thinking while attempting to balance and distribute ever shrinking fiscal resources. With this administrative attention comes the need to justify one's budget, department, teaching methodology, hiring, training, dismissal practices, and a plethora of other potential assessment opportunities. Unfortunately, accountability demands are too often made with little time allocated for adequate consideration to finding the most appropriate assessment methodology and tools so as to conduct assessment in an ethical and responsible manner. The discussion within this paper will provide an overview of issues related to the selection of assessment instruments and provide a checklist for those individuals charged with the development and/or selection of those tools.

LITERATURE REVIEW

When an individual is charged with the task of assessment, often she/he will begin immediately seeking the quickest, least expensive, and/or effective method with which to accomplish the assessment task. In addition to selecting the instrument that meets any of the above criteria, one must consider a variety of issues related to validity, reliability, and pragmatic considerations for the development and/or selection of any instrument. However, the first issue often overlooked when considering assessment is the ethical component. Unfortunately, ethical aspects of assessment are given little consideration or forgotten altogether.

Ethical Considerations

A number of ethical questions must be considered before attempting to select an assessment tool. For example, who will have access to the results of the assessment? How will the assessment data be used and with what consequences to the assessee? For example, Morley and Hulbert-Johnson (1994) suggest that the ubiquitous faculty and course evaluations are often misused to determine employment and remunerative status of faculty. While the faculty/course evaluations literature provides a range of findings relative to the reliability and validity of such instruments, there is also debate about the intended uses for these instruments. The debate is whether faculty/course evaluations are designed to be an instructional aid for faculty or are designed to be used, as currently used by administrators, to make tenure and/or salary decisions (Morley and Hulbert-Johnson, 1994).

Another ethical consideration of assessment in general is the criteria and standards to which the assessee is to be held? This issue is especially relevant with regard to ethnicity and gender. If the standard of excellence is androcentric how appropriate is it for use with females? Morreale, Moore, Taylor, Surges-Tatum, and Hulbert-Johnson (1992) address this issue with regard to assessment in the public speaking domain by providing documentation that the "The Competent Speaker" speech evaluation form has been tested for appropriate use with a variety of ethnic groups and for both genders in addition to reliability and validity testing. With regard to ethnic standards by which all will be evaluated Gomez, Ricillo, Flores, Cooper, and Starosta (1993) suggest it may not be appropriate to use only Western standards for the evaluation of intercultural communication competence. These are just a few of the issues that must be given consideration when initiating an assessment program. But these are not the only ethical aspects to

be contemplated. Much of the validity and reliability issues to be discussed below also have ethical components.

Validity Considerations

In addition to ethical aspects of assessment, validity issues must be weighed. Validity is defined as "the extent to which an empirical measure adequately reflects the *real meaning* of the concept under consideration." (Babbie, 1992, p. 132). Validity takes several forms such as face, predictive, construct, and content validity. While on the surface, one may assume that the instrument under consideration meets the validity test, all too often we may not be measuring what we think we are measuring. For example, the Bem Sex Role Inventory (Bem, 1974) may appear to have several forms of validity but under closer scrutiny, there are problems with this widely used instrument. The Bem Sex Role Inventory (BSRI), developed in the early 1970s, purports to measure the psychological construct of androgyny. Within this instrument one finds masculinity, femininity, and androgyny. What is problematic about the BSRI is that in order to define the terms of masculinity, femininity, and androgyny, Bem asked university students to provide descriptors of these terms (see Table 1). Remembering that students were providing a 1970s perception of what it means to be masculine or feminine in the US culture, the instrument may not provide a valid reflection of current understandings of what the concepts of masculinity and femininity. Therefore, the face, predictive, content and/or construct validity of the BSRI may be in question for use in conducting research in the 1990s.

Definitional issues are also relevant validity problems often encountered with instrument development and/or selection. On the surface, defining what it is one wishes to measure appears straight forward. However, if one cannot define what it is one desires to measure, predictive,

content, and construct validity will suffer. Just a few examples make clear that what appears obvious on the surface may indeed be more complex and ambiguous than first envisioned. For example, Gomez, et al. (1994) found defining intercultural communicative competence to be a challenge that has not been fully conquered. Parry (1993) also suggests when attempting to define organizational climate it becomes obvious that finding and/or developing a definition is not as obvious as one might have first thought.

TABLE 1: Masculine and Feminine Terms articulated in the BSRI (Bem, 1974)

Masculine Terms	**Feminine Terms**
• Acts as a leader	Affectionate
• Aggressive	Cheerful
• Ambitious	Childlike
• Analytical	Compassionate
• Assertive	Does not use harsh language
• Athletic	Eager to soothe hurt feelings
• Competitive	Feminine
• Defends own beliefs	Flatterable
• Dominant	Gentle
• Forceful	Gullible
• Has leadership abilities	Loves children
• Independent	Loyal
• Individualistic	Sensitive to the needs of others
• Makes decisions easily	Shy
• Masculine	Soft-Spoken
• Self-reliant	Sympathetic
• Self-sufficient	Tender
• Strong personality	Understanding
• Willing to take a stand	Warm
• Willing to take risks	Yielding

Another example of definitional confusion comes from the organizational oral communication competence (OOCC) literature. In that literature there is much debate about what

factors must be present for an individual to be considered competent. Does competence consist of knowledge (cognition), behaviors, or affect? Must all three factors be present for one to be perceived as competent? In reviewing the various conceptualization of OOCC, Jablin, et al. (1989) suggest one perspective articulates competence as performance of specific behaviors or skills while a second perspective views competence as social cognition/symbolic interaction. As a result of a comprehensive review of the extant literature, Jablin, et al (1989) developed a definition of OOCC that encompasses both the behavioral and social cognition/symbolic interaction perspectives. That definition of OOCC is

> The set of abilities, henceforth termed resources, which a communicator has available for use in the communication process. These resources are acquired via a dynamic learning process and take the form of interrelated subsets of communication skills, henceforth termed capacities, and strategic knowledge of appropriate communication behavior (Jablin, et al., 1989, p. 9).

As the above definition makes clear, effectiveness is not specifically addressed and appears not to be a requirement for perceptions of competence. So two new issues are raised relative to this definition: effectiveness and *perceptions* of competence. McCroskey (1982) argues that one may communicate competently and not achieve a specific objective. This makes sense given the transactional/interactional nature of communication (Watzlawick, Beavin, & Jackson, 1967). In other words, one may communicate competently, but one cannot control the behaviors and/or perceptions of others with whom one is communicating. Therefore, one participant in a communication event is not solely responsible for the outcomes of that communication.

Subsequent to McCroskey (1982) and Jablin et al (1989), Rubin (1990) reviewed the OOCC literature. Her review suggest that much organization literature perceives OOCC as a sum of the individual parts. Rubin (1990) summarizes the OOCC conceptualization debate thusly: "Can we dissect communication competence into constituent elements that can be taught individually, or is communication competence **basically impressions** (emphasis added) about other's behavioral predispositions." (p. 103) In other words, is competence something one possess and can be taught or is competence a matter of observer perceptions about another's behaviors?

Another example of the definitional problem is articulated by Arter (1988) in reviewing assessment of leadership and managerial skills in the educational context. Defining effective leadership proved challenging when reviewing over 40 instruments that purport to assess educational leadership and management skills.

As the above discussion indicates, defining just what is to be measured may be no easy problem to resolve. Related to that discussion is how much of any domain is the instrument measuring? Shockley-Zalabak and Hulbert-Johnson (1993; 1994) reviewed 72 assessment instruments widely used in business, industry, and government. While these instruments measure various aspects of organizational communication, few purport to measure OOCC across all organizational contexts. Users of these instruments must then take care not to assume that any one of these 72 instruments is appropriately used to measure more than that part of organizational communication specified by the instrument developer or as is evident after reviewing the instrument.

Once a clear definition has been obtained, an operational definition must be developed (Parry, 1994). In other words, what test items will actually measure the construct that has been defined. For example, from the discussion of OOCC above, do test items on any given instrument measure knowledge (cognition) , behavior or some combination of both? Which is more appropriate for the stated purpose? Unless stated by the instrument developer, users must carefully study the test for relevant validity issues.

The above discussion also exemplifies another issue related to validity. What methodology best measures what the instrument purports to measure (Parry, 1993). Should one use a self-report instrument or a trained rater/observer instrument? Self-report instruments are notoriously biased when asking individuals to recall or observe their own behavior. There is much evidence that people are not objective observers of their own behavior. For example, asking students to grade their own speeches in a public speaking class may well provide results far different than a trained observer/rater. However, self-report instruments are appropriate when asking respondents about their attitudes and values. An observer may not be able to determine attitudes and values by observing any given individual's behavior. Lau (1996) suggests that the best methodology for assessing and selecting educational institutional administrators is best accomplished using multiple observer/raters. While Lau acknowledges this is labor intensive, expensive and time-consuming, it produces the best results.

Arter (1989) also suggests that when engaged in assessment, we must consider how best to access that which we assess. In other words, can role-playing fully represent "real-world" experiences? Can recognizing something on a paper and pencil test accurately measure responses to "real-world" stimuli?

As this discussion indicates, defining that thing to be measured is a critical first step in instrument selection and/or development. Without clear definitions, operational definitions, and appropriate methodologies for accessing the thing to be measure, validity suffers.

Reliability Considerations

But validity is not the only area of concern with regard to instrument selection/development. Reliability also must be considered. Reliability is defined as how consistent an instrument behaves (Kerlinger, 1986) . Instrument developer must test for a variety of reliability issues. Does the instrument act the same way when used by various raters, i.e. inter-rater reliability? Does the instrument act the same way with similar populations? For what population is the instrument appropriately used? Can one take an instrument developed for and tested on one demographic group and apply it to another demographic group? For example, Lamude and Daniels (1990) converted Rubin's (1982) Communication Competency Assessment Instrument (CCAI) from a trained rater context instrument designed to be used at the tertiary level to a paper and pencil instrument purported to measure communication competency of organizational managers. Upon this conversion, new reliability and validity studies must be conducted to confirm the new instrument may be appropriately used with managers.

Finally, one cannot neglect the role that users of instruments play when discussing bias, validity and reliability issues. Any given instrument may have been developed with due diligence given to bias, validity, and reliability issues; they may have been fully articulated and reported in an accompanying manual. But if the user chooses to misuse the instrument, there may be serious ethical and legal consequences. "[L]et the buyer beware. If you are buying or creating an assessment instrument, the burden of proof rests with you to demonstrate to yourself, to the users

and, possibly, to a court of law that the tool is valid and reliable." (Parry, 1993, p. 42) Further, one may need to prove that as the user, you are applying and using that instrument appropriately with the intended demographic population. Therefore, the ethical aspect of assessment cannot be ignored or given little consideration.

GUIDELINES FOR INSTRUMENT SELECTION

Given the discussion above regarding appropriate selection and use of assessment tools, the following guidelines for instrument development and/or selection are provided to assist those individuals charged with assessment tasks. Guidelines are organized into test development issues and test selection issues. Each of these will be discussed in turn.

Test development issues.

There are a number of test development issues that should be addressed in an accompanying manual. Those items that ideally would be presented in the manual include: the author's purpose, construct definition, theoretical rationale, reliability and validity data, normative data and procedures, description of design procedures, item coverage, administration and scoring instructions, and interpretation information.

Purpose: The first concern of individuals selecting an instrument must be the selection of an instrument and its use as intended by the test developer. Test developers have a specific use and population in mind when developing any given instrument, therefore, it is incumbent upon test users to appropriately apply that instrument. For example, has the test been developed for use for program/needs assessment or for selection, placement, or retention of individual employees? Is the purpose and intended use clearly stated in an accompanying manual?

Construct Definition: Test developers, through their selection of measurement items and strategies, define for a specific instrument the parameters of the construct or constructs measured by that test. Of important interest, therefore, is how the author defines the construct the test purports to assess. Many instruments, for example, define communication neither at the macro nor micro level. They measure predispositions for behaviors as opposed to behaviors and assume linkage. While the measurement of predispositions in and of itself is not inappropriate, the lack of clarity about constructs has led to consistent misuse of numerous instruments purporting to measure communication behaviors. Careful test development includes theoretical rationale and related literature reviews regarding the construct as it is intended to be measured through instrumentation.

Psychometric Properties: As discussed above, there are a number of reliability, validity, and bias issues of importance for instrument development and selection. Reliability simply defined "is the accuracy or precision of a measuring instrument." (Kerlinger, 1986, p. 405). There are several types of reliability that can be relevant for a given instrument. Forms of reliability include: test-retest, split-half, alternative forms, inter-rater, and internal consistency. Test developers should address those reliabilities important for a particular instrument. Validity also must be considered. Kerlinger (1986) suggests validity is most easily understood by asking the question: "Are we measuring what we think we are measuring?" (P. 417) Again, there are several types of validity that should be considered: content (representativeness), predictive (criterion), concurrent, and construct (convergent or discriminant). Finally, there are several forms of biases of which test developers must be aware, test for, and report. For example, how has instrument development addressed cultural, gender, racial, ethnic, age, and/or international differences? How

do these issues affect the instrument validity and how should they guide appropriate instrument use?

Design Procedures and Normative Data: Developmental procedures should be clearly articulated and appropriate to the test's intended purpose, constructs measured, and supportive of generally accepted reliability and validity standards. For example, what population was used when testing the instrument? What norms have been established by that population? Does the test's author propose these norms have value-laden (good to bad) dimensions? How were the constructs and statistical limits for the values derived?

Test Administration: Complete administration, scoring and interpretation instructions should be developed by the test's author. A few of the questions that should be addressed are: How long will it take to administer the test? Are special forms needed for recording responses? Is one score adequate or are subscores appropriate? What do scores mean? What problems might be encountered during administration or interpretation of the test?

To summarize test development issues, Brown (1983) suggests that in addition to looking at the information provided by the author, one should also look for what has not been provided. "Thus not only should you adopt a critical attitude toward the information contained in a test manual, you should also pay particular attention to questions that are not answered in the manual" (Brown, 1983, p. 458).

Selection Issues

As with test development, when selecting an instrument, there are several issues to be addressed which include: user's purpose for testing, needs/expectations for the target population,

test development adequacy, construct coverage, and administration issues such as cost, format, timing, scoring, and interpretation.

User's Purpose for Testing: In selecting an instrument for testing, the user first needs to identify his/her goals for testing as well as the needs/expectations for the target population to be tested. The test user then attempts to match the needs of the test situation with selection of an instrument(s) with similar purposes and appropriate construct measurement. An issue that is often overlooked is the level of sophistication of the test user with regard to instrument selection and use. Qualification of test users "include basic knowledge of the principles of psychological measurement and the limitations of test interpretation, and the technical knowledge necessary to evaluate the claims made in the test manual...But test users must have more than general knowledge of testing procedures and practices. They must know the literature relevant to the specific tests they are using and the testing problems they may encounter." (Brown, 1983, p. 451) The implication is that the user must know the use for which the test was intended and also the population that the test appropriately targets. It is possible that a good test is not appropriate for a particular population or examiner's purpose.

Test Development Adequacy: As discussed above, psychometric and normative data should be provided and the user should have sufficient knowledge to evaluate the data for the test(s) to be used. Have all relevant reliability, validity, and biases been addressed in the manual? Further, do test items comprehensively address the construct? Are other tests needed to adequately cover the construct?

Administration, Interpretation, and Reporting: There are very important logistical and administrative considerations when considering assessment. First, although it has been suggested

that costs should not dictate the use of any given test (Brown, 1983), this is not an insignificant consideration. Not only must the test user consider the hard/electronic copy costs of the test, but time away from other assigned tasks for the target population as well as test administrator(s).

Second, are scoring and interpretation procedures clearly articulated in the manual? Can scoring and interpretation be done "in house" or must tests be submitted to the developer for scoring and interpretation? What additional expense might be incurred if the test developer provides scoring services? When interpreting the scores, what impact will the feedback have on test-takers? Test users must be sensitive to potential adverse effects on test-takers' careers and psychological health.

Finally, to whom should results be reported? How should scores be used? What are the ethical responsibilities of assessment administrators with regard to the consequences to the assessee?

CONCLUSION

Assessment tasks are increasingly consuming more time, cognitive energy, and fiscal resources of both individuals and institutions. And with this emphasis on assessment, it is incumbent upon institutions and individuals assigned those assessment tasks to conduct ethical and responsible assessment programs. The guidelines for instrument selection offered in this paper suggest that at a minimum, there are a plethora of issues that must be given careful and diligent consideration in order to conduct ethical and responsibile assessment. Responsible assessment involves time-consuming consideration of potential bias, reliability, validity, and pragmatic administration and reporting procedures. Assessment can be conducted in such a manner as to benefit both the assessee and the assessor and the sponsoring institution.

References

Arter, J. A. (1988). Assessing leadership and managerial behavior. A consumer's guide. ERIC ED300875.

Arter, J. A. (1989). Assessing communication competence in speaking and listening: A consumer's guide. ERIC ED314789.

Babbie, E. (1992). The practice of social research (6th ed.). Belmont, Ca: Wadsworth.

Bem, S. L. (1974). The measurement of psychological androgyny. Journal of Counseling and Clinical Psychology, 42, 155-162.

Brown, F. G. (1983). Principles of educational and psychological testing (3rd ed.). New York: Holt, Rinehart and Winston.

Gomez, A. M., Ricillo, S., Flores, N. L., Cooper, P., & Starosta, W. (1994). Intercultural communication competence: Developing assessment instruments. 1994 SCA Summer Conference Proceedings and Prepared Remarks. Annandale, VA: Speech Communication Association.

Jablin, F. M., Cude, R. L., Wayson, K. W., House, A., Lee, J., & Roth, N. L. (1989). Communication competence in organizations: Conceptualization and comparison across multiple levels of analysis. Paper presented at the annual convention of the International Communication Association, San Francisco, CA, May, 1989.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). Fort Worth: Holt, Rinehart and Winston.

Lamude, K. G., & Daniels, T. D. (1990). Mutual evaluations of communication competence in superior-subordinate relationships: Sex role incongruency and pro-male bias. Women's Studies in Communication, 13, (2), 39-56.

Lau, R. (1996). Applying the principles of the assessment center in the selection process of senior administrators in education. ERIC 400800.

McCroskey, J. C. (1982). Communication competence and performance: A research and pedagogical perspective. Communication Education, 31, 1-7.

Morley, D. D., & Hulbert-Johnson, R. (1994). The use and misuse of student evaluation of instruction as an assessment of instructional effectiveness: A case study of the University of Colorado. A paper presented at the Colorado Assessment Conference, Fort Collins, Colorado. March, 1994.

Morreale, S. P., Moore, M. R., Taylor, K. Phillip, Surges-Tatum, D., & Hulbert-Johnson, R. (1992). "The Competent Speaker" speech evaluation form. Annandale, Va: Speech Communication Association.

Parry, S. B. (1993). How to validate an assessment tool. Training, 30 (4), 37-42.

Rubin, R. B. (1982). Assessing speaking and listening competence at the college level: The communication competency assessment instrument. Communication Education, 31, 19-32.

Rubin, R. B. (1990). Communication competence. In G. M. Phillips & J. T. Wood (Eds.) Speech communication: Essays to commemorate the 75th anniversary of the Speech Communication Association. Carbondale, IL: Southern Illinois University.

Shockley-Zalabak, P., & Hulbert-Johnson, R. (1993, November). Oral communication competency and its assessment in organizations: Approaches, instruments, and methods. A paper presented at the Speech Communication Association Convention, Miami Beach, Florida.

Shockley-Zalabak, P., & Hulbert-Johnson, R. (1994). Organizational communication assessment: Frameworks and methods. In W. G. Christ (Ed.) Assessing communication education: A handbook for media, speech, and theatre educators. Hillsdale, NJ: Lawrence Erlbaum.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). Pragmatics of human communication: A study of interpersonal patterns, pathologies, and paradoxes. New York: W. W. Norton.